Eötvös Loránd Tudományegyetem
Bölcsészettudományi Kar

Doktori Disszertáció Tézisei

Recski Gábor

Számítógépes módszerek a szemantikában

Nyelvtudományi Doktori Iskola
Tolcsvai Nagy Gábor MHAS

Elméleti Nyelvészet Doktori Program
Bánréti Zoltán CSc.

A bizottság tagjai:
Kiefer Ferenc MHAS (elnök)
Rebrus Péter PhD. (titkár)
Vincze Veronika PhD.
Alberti Gábor DSc.
Komlósy András CSc.

Témavezető:
Kornai András DSc.

Budapest, 2016

Bevezetés

Az értekezés bemutatja a `4lang` szoftvercsomagot, mely automatikusan állít elő `4lang`-stílusú szemantikai reprezentációkat nyers angol és magyar szövegből, egynyelvű szótárak feldolgozása által pedig új definíciós gráfokat is épít. A disszertáció bemutat egy-egy, szó- ill. mondatpárok szemantikai hasonlóságát mérő rendszert is, melyek `4lang` reprezentációkat is használnak, egyikük pedig a `4lang` definíciók segítségével éri el a jelenleg ismert legjobb eredményt a SimLex adaton, mely a szóhasonlóságot mérő rendszerek összehasonlításának legnépszerűbb terepe. Ez a füzet felsorolja az értekezés téziseit, rövidesen ismerteti a teljes mű felépítését, tisztázza a társszerzők szerepét minden alrendszer esetében, online elérhetőséget ad valamennyi szoftverhez, végül pedig röviden összefoglalja a tézisek tartalmát.

Tézisek

Az értekezésben ismertetett legfőbb új eredmények az alábbiak:

(T1) Az angol és magyar nyers szövegből `4lang` szemantikai reprezentációt előállító `text_to_4lang` eszköz

(T2) Az angol és magyar egynyelvű szótárak bejegyzéseiből `4lang` definíciós gráfokat előállító `dict_to_4lang` eszköz

(T3) Angol mondatpárok hasonlóságának mérésére szolgáló, a `dict_to_4lang` által épített definíciós gráfokat is felhasználó rendszer

(T4) Az angol szópárok hasonlóságát a jelenleg ismert legnagyobb pontossággal közelítő, a `4lang` gráfokból kinyert jegyeket is hasznosító eljárás

A disszertáció felépítése

A 2. fejezet rövid összefoglalását adja a szójelentés néhány jelentős elméletének, különös tekintettel azok nyelvtechnológiai alkalmazásaira. A 3. fejezet áttekinti a jelentésmodellezésre szolgáló `4lang` formalizmust, de részletesen nem tárgyalja annak elméleti hátterét, mivel az nem a jelen értekezés tézise, hanem féltucat kutató közös munkájának eredménye (Kornai et al., 2015). A 4. fejezet mutatja be a nyers szöveget `4lang` reprezentációra képező `dep_to_4lang` eszközt, az 5. fejezet pedig ennek egy kiemelten fontos alkalmazásával, az egynyelvű szótárak definícióiból fogalmi szótárakat építő `dict_to_4lang` rendszerrel foglalkozik. A 6. fejezet a `4lang` szoftvercsomag további alkalmazásait mutatja be, így többek között egy-egy, angol mondat- ill. szópárok szemantikai hasonlóságát mérő, gépi tanuláson alapuló rendszert, melyek közül utóbbi a ma legelterjedtebb `SimLex` adaton kiértékelve a feladatra ismert legpontosabb algoritmusnak bizonyul (Recski & Ács, 2015). A fejezet végül röviden megemlít egy kísérleti keretrendszert (Nemeskey et al., 2013), mely `4lang` reprezentációk segítségével old meg egyszerű gépi megértési feladatokat. A 7. fejezet bemutatja a kb. 3000 soros `4lang` kódbázist és röviden leírja az egyes modulok felépítését. Ez a fejezet egyben

Rendszer	Forráskód	Fő publikáció
`4lang`	`github.com/kornai/4lang`	(Recski, 2016)
`pymachine`	`github.com/kornai/pymachine`	
`semeval`	`github.com/juditacs/semeval`	(Recski & Ács, 2015)
`4lang`	`github.com/recski/wordsim`	(Recski et al., 2016)

1. táblázat. Az értekezésben bemutatott szoftverek

a szoftvercsomag dokumentációjaként is szolgál. Az értekezés utolsó, 8. fejezete a `4lang` rendszer jövőbeli alkalmazására vonatkozó terveinkről szól.

Együttműködések

A 3. fejezetben bemutatott `4lang` elmélet az MTA Matematikai Nyelvészeti Kutatócsoport volt és jelenlegi tagjainak (Ács Judit, Borbély Gábor, Kornai András, Makrai Márton, Nemeskey Dávid, Pajkossy Katalin, Zséder Attila) közös eredménye. A 4. és 5. fejezetekben ismertetett rendszerek a szerző önálló fejlesztései, az alábbi két kivétellel: a gráfexpanziót végző függvény (5.3. fejezet) Borbély Gáborral közös fejlesztés, a Collins szótárt feldolgozó modul pedig Bolevácz Attila munkája. A 6.1. fejezetben bemutatott Semeval rendszer Ács Judittal, a 6.2-ben tárgyalt `wordsim` rendszer Iklódi Eszterrel (BME Automatizálási és Alkalmazott Informatikai Tanszék) közös munka eredménye, utóbbinak gépi tanulási komponense Pajkossy Katalin fejlesztései alapján készült. A 6.3. fejezetben említett kisérleti rendszer implementációjában a szerzőn kívül Nemeskey Dávid és Zséder Attila vett részt.

Szoftver

A disszertációban bemutatott valamennyi szoftver az MIT licenc szerint szabadon letölthető, az egyes hivatkozásokat a 1. táblázat tartalmazza. A `text_to_4lang` (T1) és `dict_to_4lang` (T2) eszközök a `4lang` könyvtár részei, egyes függőségek a `pymachine` csomagban találhatók. A `4lang` kódbázis az értekezés leadásakor aktuális változatát a `recski_thesis` branch rögzíti. A mondathasonlóságot mérő rendszer (T3) forrása a `semeval` repozitóriumban található, a szóhasonlóságot mérő rendszert (T4) a `wordsim` csomag tartalmazza. Ezen eszközök valamennyi külső függősége nyílt forráskódú és különböző licencek alatt szabadon letölthető.

(T1) Az angol és magyar nyers szövegből `4lang` szemantikai reprezentációt előállító `text_to_4lang` eszköz

A `text_to_4lang` modul a `4lang` könyvtár részeként nyers szöveget képez `4lang` jelentésreprezentációkra oly módon, hogy a bemeneten sztenderd függőségi elemzőket (dependency parser) futtat, az azok kimenetében szereplő relációkat pedig `4lang` részgráfoknak felelteti meg. Egyes speciális szerkezeteket (pl. koordináció, kopuláris mondatok) a dependencia-relációkat utófeldolgozó ad-hoc szabályok kezelik. Noha a végső kimenet minőségének jelenleg határt szab, hogy csak a külső mondattani elemzők által is kezelt mintákat képesek feldolgozni, a `text_to_4lang` egyszerűbb mondatok és frázisok esetében mégis nagy pontosságú szemantikai reprezentációkat készít, melyek alapjául szolgálnak a disszertációban bemutatott valamennyi alkalmazásnak.

(T2) Az angol és magyar egynyelvű szótárak bejegyzéseiből `4lang` definíciós gráfokat előállító `dict_to_4lang` eszköz

A `4lang` könyvtár `dict_to_4lang` modulja egynyelvű szótárak bejegyzései alapján `4lang` definíciós gráfokat épít. A `dict_to_4lang` a `text_to_4lang` eszközre támaszkodik, de annak funkcionalitását több, az egyes bemenettípusokra (3 angol és 2 magyar szótár) specifikus feldolgozási lépéssel terjeszti ki. A kimeneten végzett kézi kiértékelés alapján a `dict_to_4lang` magas pontosságú reprezentációkat készít az angol szavak több mint 80, a magyar szavak több mint 60 százaléka esetében. Mivel a definíciós gráfok így csaknem valamennyi angol és magyar szóhoz elkészíthetők, lehetővé válik a gráfok *expanziója*, melynek során az egyes gráfok kibővülnek az azokban szereplő fogalmak definícióival, így minden szó jelentése visszavezethető alapfogalmak tetszőlegesen szűk halmazára.

(T3) Angol mondatpárok hasonlóságának mérésére szolgáló, a `dict_to_4lang` által épített definíciós gráfokat is felhasználó rendszer

Az angol mondatpárok szemnatikai hasonlóságának mérésére szolgáló eszközünk alapja egy korábban több `Semeval` versenyen sikerrel alkalmazott architektúra, mely a mondatok hasonlóságát a szóhasonlóságra vezeti vissza. Rendszerünk képes több szóhasonlósági metrikát kombinálni, így a hagyományos módszereket kiegészíti `4lang` definciós gráfok párjai fölött definiált hasonlósági jegyekkel is, melyek egyrészt a definíciók közös részgráfjait jellemzik, másrészt néhány, a szemantikai hasonlóságra jellemző szabályszerűséget tárnak fel. Legjobban teljesítő rendszerünk a 2015-ös versenyadaton 0.78-as Pearson-korrelációt ért el, ezzel a 78 résztvevő rendszer közül a 12. helyen végzett.

(T4) Az angol szópárok hasonlóságát a jelenleg ismert legnagyobb pontossággal közelítő, a `4lang` gráfokból kinyert jegyeket is hasznosító eljárás

Az angol szópárok jelentésbeli hasonlóságát mérő `wordsim` rendszer éri el a humán annotátorokkal mért legmagasabb korrelációt a kiértékelésre leggyakrabban használt `SimLex` adaton. Modellünk sikerrel kombinál több disztibucionális jelentésmodellt, a WordNet-ből kinyert jegyeket, valamint a `4lang`-gráfok felett definiált jegyeket (utóbbiaknak egy részét a (T3)-ban bemutatott rendszer is használta). A kizárólag vektor-alapú módszereket kombináló konfigurációnk is jobban teljesít, mint valamennyi korábban publikált rendszer, azonban azt is megmutatjuk, hogy a `4lang`-alapú jegyek további jelentős, a WordNet-jegyeknél jóval nagyobb javulást eredményeznek. A legmagasabb, 0.76-os korrelációs pontszám jóval meghaladja a humán annotátorok között átlagosan mért 0.67-es értéket, továbbá megközelíti az egy-egy annotátor a többi annotátor pontszámainak átlagával mért korrelációját is (0.78), mely arra enged következtetni, hogy rendszerünk ezen az adathalmazon megközelítette az emberi teljesítményt.

Hivatkozások

Kornai, A., Ács, J., Makrai, M., Nemeskey, D. M., Pajkossy, K., & Recski, G. (2015). Competence in lexical semantics. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (*SEM 2015)* (pp. 165–175). Denver, Colorado: Association for Computational Linguistics.

Nemeskey, D., Recski, G., Makrai, M., Zséder, A., & Kornai, A. (2013). Spreading activation in language understanding. In *Proceedings of the 9th International Conference on Computer Science and Information Technologies (CSIT 2013)* (pp. 140–143). Yerevan, Armenia: Springer.

Recski, G. (2016). Building concept graphs from monolingual dictionary entries. In N. Calzolari et al. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* Portorož, Slovenia: European Language Resources Association (ELRA).

Recski, G., & Ács, J. (2015). MathLingBudapest: Concept networks for semantic similarity. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)* (pp. 543–547). Denver, Colorado: Association for Computational Linguistics.

Recski, G., Iklódi, E., Pajkossy, K., & Kornai, A. (2016). Measuring semantic similarity of words using concept networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP* (pp. 193–200). Berlin, Germany: Association for Computational Linguistics.